

EXPANSION-MODIFICATION SYSTEM: STATIONARY MEASURES AND ASYMPTOTIC SCALING

R. SALGADO-GARCÍA AND E. UGALDE

ABSTRACT. This work is devoted to the study of the stationary measures of the expansion-modification model. We prove all initial distributions converge towards a unique stationary measure exhibiting decay of correlations. We also develop an argument towards the proof of an asymptotic scaling behavior for the correlation function, which allows us to give a closed expression for the scaling exponent as a function of the mutation probability. Finally, we prove the validity of the asymptotic scaling behavior and the corresponding expression for scaling exponents, for low mutation probability.

1. INTRODUCTION.

Since the advent of DNA sequencing techniques the problem of unveiling the information contained in nucleotide sequences has become a present-day challenge for geneticists. A specific issue which has received attention is the problem of genome evolution. From the theoretical point of view, having a “good” model for the genome evolution is fundamental to the phylogenetic reconstruction, a fast-growing field with numerous applications in a broad range of biological areas [7]. For this purpose, several models have been introduced (such as n -step Markov chains or hidden Markov chains, among others [6, 8, 13, 14]) to describe the evolution of nucleotide sequences as well as the patterns and correlations occurring in the genome. In this paper we are concerned with the model proposed by W. Li [3] which consists of a sequence (or chain) of symbols that evolve according to a given discrete-time stochastic dynamics. Such a dynamics captures the essential processes which are assumed to be responsible of the genome evolution: the random expansion and modification of symbols, rules that gave the name to the system of *expansion-modification model*. The latter was originally introduced as a simple model exhibiting some spatial scaling properties, a behavior which is ubiquitous to several phenomena found in nature [3]. Subsequently this was used to understand the scaling properties and the long-range correlations found in real DNA sequences [2, 4, 5, 6, 12]. Recently the expansion-modification system has also been used to investigate the universality of the rank-ordering distributions [1, 10].

The model we will deal with in this paper can be described as follows. Consider the random substitution ¹

$$x \mapsto \begin{cases} \bar{x} & \text{with probability } p, \\ xx & \text{with probability } 1 - p, \end{cases}$$

Date: February 14, 2012.

¹The upper bar indicates “negation”, *i.e.*, $\bar{0} = 1$ and $\bar{1} = 0$

in the binary set $\{0, 1\}$, and extend it coordinate-wise to the set $\{0, 1\}^+$ of finite binary strings. Starting at time zero with a seed in $\mathbf{x}^0 \in \{0, 1\}^+$, and iterating the above substitution, we obtain a sequence

$$\mathbf{x}^0 \mapsto \mathbf{x}^1 \mapsto \cdots \mapsto \mathbf{x}^n \mapsto \cdots$$

of finite strings of non-decreasing length. Since the applied substitution is a random map, the sequence we obtain by successive iterations is a random sequence which is nevertheless supposed to converge, in a certain statistical sense, to a random string \mathbf{x}^∞ . It is easy to see that the probability of having a finite string after infinitely many iterations is zero, it is therefore more convenient to studying the evolution of infinite strings under the infinite extension of the above substitution. In this framework we can rigorously define the asymptotic regime of the expansion-modification process and single out some of its salient characteristics. In this work we present a mathematical study of the asymptotic regime of the expansion-modification dynamics and present a rigorous proof of the scaling behavior of the correlation function, which allows us to determine a closed form for the scaling exponent.

The paper is organized as follows: in Section 2 we set up the mathematical framework where the expansion-modification system is rigorously defined, then, in Section 3 we prove the existence of a unique stationary distribution towards which the dynamics converges. In Section 4 we prove that the unique stationary measure exhibits decay of correlations. In Section 5 we prove the asymptotic scaling of the correlation function, and we compute a closed expression for the corresponding scaling exponent. We finish the paper with some final remarks and comments.

This work was partially supported by CONACyT-Mexico via the grant No. 129072 and SEP-Mexico through the PIFI program. The authors are in debt to G. Salazar for his careful reading of the manuscript and the resulting suggestions.

2. THE EXPANSION-MODIFICATION DYNAMICS.

2.1. We start with some notations. Let $X = \{0, 1\}^{\mathbb{N}_0}$, endowed X with the σ -algebra generated by the cylinder sets. Elements of X will be denoted by boldface characters like $\mathbf{x} = x_0x_1\cdots$ and $\mathbf{y} = y_0y_1\cdots$, where $x_i, y_i \in \{0, 1\}$. Finite sequences of symbols, also called *words*, will be also denoted by boldfaced letters while their size will be denoted by $|\cdot|$, i.e., for $\mathbf{v} \in \{0, 1\}^k$ we have $|\mathbf{v}| = k$. A word $\mathbf{v} \in \{0, 1\}^k$ occurs as prefix of $\mathbf{x} \in X$, which we denote by $\mathbf{v} \sqsubseteq \mathbf{x}$, if $\mathbf{v} = x_0x_1\cdots x_{k-1}$. We will also use this notation when $\mathbf{x} \in X$ is replaced by a finite word. Given a configuration $\mathbf{x} \in X$ and integers $0 \leq i < j$, with \mathbf{x}_i^j we denote the word $x_ix_{i+1}\cdots x_j$ occurring in \mathbf{x} . Product of words will be understood as concatenation: given two words $\mathbf{v} \in \{0, 1\}^k$ and $\mathbf{w} \in \{0, 1\}^l$ we will denote the word \mathbf{u} of size $k + l$, with prefix \mathbf{v} and suffix \mathbf{w} , by \mathbf{vw} ; that is $\mathbf{u}_0^{k-1} = \mathbf{v}$ and $\mathbf{u}_k^{k+l-1} = \mathbf{w}$. Consider $S = \{e, m\}^{\mathbb{N}_0}$, where the symbols e and m stand for expansion and modification respectively. The space S , which we will refer to as the *space of substitutions*, is endowed with the σ -algebra generated by the cylinder sets as well. We will use the same convention to denote the elements of S , words and concatenation of words, as for the symbolic space X .

2.2. Let us now define the *local substitutions* $e, m : \{0, 1\} \rightarrow \{0, 1\}^+ := \cup_{n=1}^{\infty} \{0, 1\}^n$, which are given by

$$\begin{aligned} e(x) &= xx, \\ m(x) &= \bar{x}. \end{aligned}$$

A sequence $\mathbf{s} \in S$ of local substitutions defines the *global substitution* $\mathbf{s} : X \rightarrow X$ by means of the rule

$$\mathbf{s}(\mathbf{x}) = \prod_{i \in \mathbb{N}_0} s_i(x_i).$$

Here \prod stands for concatenation of words. Notice that \mathbf{s} replaces the i -th symbol of \mathbf{x} according to the i -th local substitution, *i.e.*, if $s_i = e$ then x_i is expanded, otherwise x_i is modified.

2.3. The *expansion-modification dynamics* is a random dynamical system whose orbits depend on an initial condition and a choice of global substitutions to be applied to that initial condition. To be more precise, an initial condition $\mathbf{x} \in X$ and a sequence $\mathbf{s}^0, \mathbf{s}^1, \mathbf{s}^2, \dots$ of configuration in S , define the orbit $\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2, \dots$ in X with $\mathbf{x}^0 := \mathbf{x}$ and for each $t > 0$, $\mathbf{x}^{t+1} = \mathbf{s}^t(\mathbf{x}^t)$. The choice of the sequences of global substitutions to be applied is determined by a probability measure on S . At each time step we randomly choose a configuration $\mathbf{s} \in S$ according to that measure, and then we apply the corresponding global substitution. The measure according to which we select the sequences of substitutions is taken from the family of Bernoulli measures $\{\nu_p : p \in [0, 1]\}$ defined as follows: for a given $p \in [0, 1]$, ν_p is the product measure such that $\nu_p[m] = p$ and $\nu_p[e] = 1 - p$, *i.e.*, ν_p corresponds to a random sequence of modifications and expansions which are selected independently and uniformly, with probability p for the first and $1 - p$ for the second. Below we will refer to p as the *mutation probability*. The probability of events involving a finite number of coordinates depends only on the measures of cylinder sets in S and in X . We say that the system has an *asymptotic behavior* if regardless of the measures describing the initial distribution, the time- t distribution converges in the weak* sense as t goes to infinity, *i.e.*, the probability of all events involving a finite number of coordinates converges as the time goes to infinity.

2.4. Let us assume that the time- t configurations, \mathbf{x}^t , are distributed according to the measure μ^t on X . The distribution μ^{t+1} of time- $(t+1)$ configurations is completely determined by ν_p and μ^t according to the following expression:

$$(1) \quad \mu^{t+1}\{(\mathbf{x}^{t+1})_0^\ell = \mathbf{a}\} = \sum_{\mathbf{c} \in \{e, m\}^{\ell+1}} \sum_{\substack{\mathbf{b} \in \{0, 1\}^{\ell+1} \\ \mathbf{a} \sqsubseteq \prod_{i=0}^{\ell} s_i(b_i)}} \mu^t\{(\mathbf{x}^t)_0^\ell = \mathbf{b}\} \nu_p\{(\mathbf{s}^t)_0^\ell = \mathbf{c}\},$$

for each $\ell \in \mathbb{N}_0$ and $\mathbf{a} \in \{0, 1\}^\ell$. As mentioned before, $\mathbf{a} \sqsubseteq \mathbf{b}$ means that the word \mathbf{a} occurs as a suffix of the word \mathbf{b} . Hence, the evolution of the length- ℓ marginal is nothing but a Markov chain. Indeed, considering the length- ℓ marginal of a measure μ as a probability vector of dimension $2^{\ell+1}$, the length- ℓ marginal μ_ℓ^t , of the time- t distribution is given by matrix product $\mu_\ell^t = \mu_\ell^0 M_\ell^t$, where $M_\ell : \{0, 1\}^{\ell+1} \times \{0, 1\}^{\ell+1} \rightarrow [0, 1]$ is the $2^{\ell+1} \times 2^{\ell+1}$ -stochastic matrix given by

$$M_\ell(\mathbf{a}, \mathbf{b}) = \sum_{\mathbf{c} \in \{e, m\}^{\ell+1}} \sum_{\substack{\mathbf{b} \in \{0, 1\}^{\ell+1} \\ \mathbf{a} \sqsubseteq \prod_{i=0}^{\ell} s_i(b_i)}} \nu_p[\mathbf{c}].$$

3. THE ASYMPTOTIC DISTRIBUTION.

Let us assume for the moment that for each $\ell \in \mathbb{N}_0$, the stochastic matrix M_ℓ is primitive. Then, the Perron–Frobenius Theorem ensures the existence of a unique probability vector $\mu_\ell : \{0, 1\}^{\ell+1} \rightarrow (0, 1]$ such that

$$\mu_\ell = \mu_\ell M_\ell \quad \text{and} \quad \lim_{t \rightarrow \infty} \mu_\ell^0 M_\ell^t = \mu_\ell,$$

for each probability vector $\mu_\ell^0 : \{0, 1\}^{\ell+1} \rightarrow [0, 1]$. Hence, for any measure μ^0 specifying the distribution of the initial conditions, for each $\ell \in \mathbb{N}_0$, and for all $\mathbf{a} \in \{0, 1\}^{\ell+1}$ we have

$$(2) \quad \lim_{t \rightarrow \infty} \mu^t[\mathbf{a}] = \mu_\ell(\mathbf{a}).$$

If in addition the probability vectors μ_ℓ satisfy the compatibility condition

$$(3) \quad \sum_{x \in \{0, 1\}} \mu_{\ell+1}(\mathbf{a}x) = \mu_\ell(\mathbf{a}),$$

for each $\ell \in \mathbb{N}_0$ and $\mathbf{a} \in \{0, 1\}^{\ell+1}$, then Kolmogorov’s Consistency Theorem implies the existence of a measure μ on X such that $\mu[\mathbf{a}] = \mu_\ell(\mathbf{a})$ for each $\ell \in \mathbb{N}_0$ and $\mathbf{a} \in \{0, 1\}^{\ell+1}$. Finally, Equation (2) ensures the convergence of μ^t towards μ in the weak* sense.

The primitivity of M_ℓ is a consequence of the following argument. As we prove in Appendix A, for each pair of words $\mathbf{a}, \mathbf{b} \in \{0, 1\}^{\ell+1}$, there exists a sequence of substitutions such that applied to \mathbf{a} produces a word having \mathbf{b} as prefix. Now, since all words in $\{e, m\}^{\ell+1}$ have positive probability, then the previous claim implies that $M_\ell^n(\mathbf{a}, \mathbf{b}) > 0$ for some $n > 0$, which proves that M_ℓ^n is irreducible. Now, since the world $00 \cdots 0$ occurs as the prefix of $e(0)e(0) \cdots e(0)$, then $M_\ell(00 \cdots 0, 00 \cdots 0) > 0$, which implies that M_ℓ is aperiodic, therefore M_ℓ is primitive.

Now, the compatibility condition (3) is inherited from the analogous compatibility condition satisfied by the marginals μ_ℓ^t at each time $t \in \mathbb{N}_0$. Indeed, for $t = 0$ we obviously have

$$\sum_{x \in \{0, 1\}} \mu_{\ell+1}^0(\mathbf{a}x) := \sum_{x \in \{0, 1\}} \mu^0[\mathbf{a}x] = \mu^0 \left(\bigsqcup_{x \in \{0, 1\}} [\mathbf{a}x] \right) = \mu[\mathbf{a}] =: \mu_\ell^0(\mathbf{a}),$$

for each $\ell \in \mathbb{N}_0$ and $\mathbf{a} \in \{0, 1\}^{\ell+1}$. Here \sqcup stands for the disjoint union. Now, from Equation (1) it follows that

$$\begin{aligned} \sum_{x \in \{0, 1\}} \mu_{\ell+1}^{t+1}(\mathbf{a}x) &= \sum_{x \in \{0, 1\}} \sum_{\mathbf{s} \in \{e, m\}^{\ell+2}} \nu_p[\mathbf{s}] \left(\sum_{\substack{\mathbf{b} \in \{0, 1\}^{\ell+2} \\ \mathbf{a}x \sqsubseteq \prod_{i=0}^{\ell+1} s_i(b_i)}} \mu^t[\mathbf{b}] \right) \\ &= \sum_{\mathbf{s} \in \{e, m\}^{\ell+2}} \nu_p[\mathbf{s}] \mu^t \left(\bigsqcup_{x \in \{0, 1\}} \bigsqcup_{\substack{\mathbf{b} \in \{0, 1\}^{\ell+2} \\ \mathbf{a}x \sqsubseteq \prod_{i=0}^{\ell+1} s_i(b_i)}} [\mathbf{b}] \right) \\ &= \sum_{\mathbf{s} \in \{e, m\}^{\ell+2}} \nu_p[\mathbf{s}] \mu^t \left(\bigsqcup_{\substack{\mathbf{b} \in \{0, 1\}^{\ell+2} \\ \mathbf{a} \sqsubseteq \prod_{i=0}^{\ell+1} s_i(b_i)}} [\mathbf{b}] \right). \end{aligned}$$

Since $|\mathbf{a}| = \ell + 1$, the statement $\mathbf{a} \sqsubseteq \prod_{i=0}^{\ell+1} s_i(b_i)$ is equivalent to $\mathbf{a} \sqsubseteq \prod_{i=0}^{\ell} s_i(b_i)$, and we have

$$\begin{aligned} \sum_{x \in \{0, 1\}} \mu_{\ell+1}^{t+1}(\mathbf{a}x) &= \sum_{\mathbf{s} \in \{e, m\}^{\ell+2}} \nu_p[\mathbf{s}] \mu^t \left(\bigsqcup_{\substack{\mathbf{b} \in \{0, 1\}^{\ell} \\ \mathbf{a} \sqsubseteq \prod_{i=0}^{\ell} s_i(b_i)}} [\mathbf{b}] \right) \\ &= \sum_{\mathbf{s} \in \{e, m\}^{\ell+1}} \left(\sum_{\substack{\mathbf{b} \in \{0, 1\}^{\ell+1} \\ \mathbf{a} \sqsubseteq \prod_{i=0}^{\ell} s_i(b_i)}} \mu^t[\mathbf{b}] \right) \sum_{\rho \in \{e, m\}} \nu_p[\mathbf{s}\rho] \\ &= \sum_{\mathbf{s} \in \{e, m\}^{\ell}} \left(\sum_{\substack{\mathbf{b} \in \{0, 1\}^{\ell+1} \\ \mathbf{a} \sqsubseteq \prod_{i=0}^{\ell} s_i(b_i)}} \mu^t[\mathbf{b}] \right) = (\mu_{\ell}^t M_{\ell})(\mathbf{a}) := \mu_{\ell}^{t+1}(\mathbf{a}) \end{aligned}$$

for each $\ell \in \mathbb{N}_0$ and $\mathbf{a} \in \{0, 1\}^{\ell+1}$. The compatibility condition (3) follows by taking the limit $t \rightarrow \infty$ on both sides of the equation.

The above arguments prove of the following:

Theorem 1. *For each $p \in (0, 1)$ there exists a unique measure μ_p on X which is invariant under the expansion–modification dynamics. Furthermore, starting from any measure μ^0 determining the distribution of the initial conditions, the measure μ^t , corresponding to the distribution at time t , converges in the weak* sense to μ_p .*

The above discussion applies to any random global substitution defined by a finite collection of substitutions on a finite alphabet. The existence of an asymptotic behavior depends only on the fact that the stochastic matrices governing the evolution of the marginals are primitive.

4. DECAY OF CORRELATIONS.

The *two-sites correlation function*, $C_p : \mathbb{N}_0 \rightarrow \mathbb{R}$, is given by

$$C_p(n) := \int_X \mathbf{x}_0 \mathbf{x}_n d\mu_p(\mathbf{x}) - \left(\int_X \mathbf{x}_0 d\mu_p(\mathbf{x}) \right) \left(\int_X \mathbf{x}_n d\mu_p(\mathbf{x}) \right),$$

where \mathbf{x}_n denotes, as usual, the projection of \mathbf{x} on the n -th coordinate. Since the expansion-modification dynamics is invariant under the flip $\mathbf{x}_n \leftrightarrow \overline{\mathbf{x}_n}$ for each $n \in \mathbb{N}_0$, then the stationary measure μ_p has to satisfy $\mu_p\{\mathbf{x}_n = 1\} = \mu_p\{\mathbf{x}_n = 0\} = 1/2$ for all $n \in \mathbb{N}_0$. Therefore

$$C_p(n) := \int_X \mathbf{x}_0 \mathbf{x}_n d\mu_p(\mathbf{x}) - 1/4 \equiv \mu_p\{\mathbf{x}_0 = \mathbf{x}_n = 1\} - 1/4.$$

We will prove that μ_p has decay of correlations, *i.e.*, that $\lim_{n \rightarrow \infty} C_p(n) = 0$, for each $p \in (0, 1)$.

Since μ_p is flip-invariant, then $\mu_p\{\mathbf{x}_0 = \mathbf{x}_n = 1\} = \mu_p\{\mathbf{x}_0 = \mathbf{x}_n = 0\}$ and $\mu_p\{\mathbf{x}_0 = 0 \neq \mathbf{x}_n = 1\} = \mu_p\{\mathbf{x}_0 = 1 \neq \mathbf{x}_n = 0\}$, for each $n \in \mathbb{N}_0$, hence

$$C_p(n) := \frac{1}{2} (\mu_p\{\mathbf{x}_0 = \mathbf{x}_n\} - 1/2) = \frac{1}{4} (\mu_p\{\mathbf{x}_0 = \mathbf{x}_n\} - \mu_p\{\mathbf{x}_0 \neq \mathbf{x}_n\}).$$

Now, since μ_p is invariant under the expansion-modification dynamics, then

$$\begin{aligned} \mu_p\{\mathbf{x}_0 = \mathbf{x}_n\} &= \sum_{k=\lceil n/2 \rceil}^n \mu_p\{\mathbf{x}_0 = \mathbf{x}_k\} \nu_p\{\mathbf{s}_0 = \mathbf{s}_k, \ell(\mathbf{s}_0^k) = n+1\} + \\ &\quad \sum_{k=\lceil n/2 \rceil}^n \mu_p\{\mathbf{x}_0 = \mathbf{x}_k\} \nu_p\{\mathbf{s}_0 = \mathbf{s}_k = e, \ell(\mathbf{s}_0^k) = n+2\} + \\ &\quad \sum_{k=\lceil n/2 \rceil}^n \mu_p\{\mathbf{x}_0 \neq \mathbf{x}_k\} \nu_p\{\mathbf{s}_0 \neq \mathbf{s}_k, \ell(\mathbf{s}_0^k) = n+1\} + \\ &\quad \sum_{k=\lceil n/2 \rceil}^n \mu_p\{\mathbf{x}_0 \neq \mathbf{x}_k\} \nu_p\{\mathbf{s}_0 \neq \mathbf{s}_k = e, \ell(\mathbf{s}_0^k) = n+2\}, \end{aligned}$$

and similarly for $\mu_p\{\mathbf{x}_0 \neq \mathbf{x}_n\}$. Here and below we denote by $\ell(\mathbf{s}_0^k)$ the length of the words obtained by applying the substitution \mathbf{s}_0^k . From the previous equation and its analogous for $\mu_p\{\mathbf{x}_0 \neq \mathbf{x}_n\}$, it follows that

$$(4) \quad C_p(n) = \sum_{k=\lceil n/2 \rceil}^n C_p(k) (f(p) \nu_p(k, n) + g(p) \nu_p(k, n-1) + h(p) \nu_p(k, n-2)),$$

where

$$\begin{aligned} f(p) &:= p(2p-1), \\ g(p) &:= (1-p)(1-3p), \\ h(p) &:= (1-p)^2, \end{aligned}$$

and for each $k, n \in \mathbb{N}$

$$(5) \quad \nu_p(k, n) := \nu_p\{\ell(\mathbf{s}_1^{k-1}) = n-1\} \equiv \binom{k-1}{n-k} (1-p)^{n-k} p^{2k-n-1}.$$

From this point on we will use the notation $S_p(n) := \sum_{k=\lceil n/2 \rceil}^n \nu_p(k, n)$. It follows, from a straightforward computation, that $S_p(n+1) = pS_p(n) + (1-p)S_p(n-1)$. This recursion, starting from $S_p(0) = 0$ and $S_p(1) = 1$, gives

$$(6) \quad S_p(n) = \frac{1 - (p-1)^n}{2-p}.$$

Now, since for $0 \leq p \leq 1/3$ we have $f(p) \leq 0 \leq g(p), h(p)$, then, by taking absolute values on both sides of (4) we obtain

$$|C_p(n)| \leq \max_{n/2 \leq k \leq n} |C_p(k)| (-f(p) S_p(n) + g(p) S_p(n-1) + h(p) S_p(n-2)).$$

From this and (6) we obtain

$$(7) \quad |C_p(n)| \leq \max_{n/2 \leq k \leq n} |C_p(k)| ((1-2p) + 2p(1-p)^n).$$

For $1/3 \leq p \leq 1/2$ we have $f(p), g(p) \leq 0 \leq h(p)$, therefore

$$(8) \quad \begin{aligned} |C_p(n)| &\leq \max_{n/2 \leq k \leq n} |C_p(k)| \max(-f(p) S_p(n) - g(p) S_p(n-1) + h(p) S_p(n-2)) \\ &\leq \max_{n/2 \leq k \leq n} |C_p(k)| \left(\frac{p(3-4p)}{2-p} + \frac{2(1-p)^n(1-p-p^2)}{2-p} \right). \end{aligned}$$

Finally, for $1/2 \leq p \leq 1$ we have $g(p) \leq 0 \leq f(p), h(p)$, and then

$$(9) \quad \begin{aligned} |C_p(n)| &\leq \max_{n/2 \leq k \leq n} |C_p(k)| \max(-g(p) S_p(n-1) + f(p) S_p(n) + h(p) S_p(n-2)) \\ &\leq \max_{n/2 \leq k \leq n} |C_p(k)| \left(\frac{p}{2-p} + \frac{2p(2p-1)(1-p)^n}{2-p} \right). \end{aligned}$$

All of the inequalities (7), (8) and (9) have the form

$$|C_p(n)| \leq \max_{n/2 \leq k \leq n} |C_p(k)| (\alpha(p) + \epsilon_p(n)),$$

with $\alpha(p) \in (0, 1)$ and $\lim_{n \rightarrow \infty} \epsilon_p(n) = 0$. Taking the limsup in both sides of the inequality we obtain

$$\limsup_{n \rightarrow \infty} |C_p(n)| \leq \alpha(p) \limsup_{n \rightarrow \infty} \max_{n/2 \leq k \leq n} |C_p(k)| = \alpha(p) \limsup_{n \rightarrow \infty} |C_p(n)|,$$

which implies $\lim_{n \rightarrow \infty} C_p(n) = 0$.

In this way we have proved the following:

Theorem 2. *For each $p \in (0, 1)$, the stationary measure μ_p has decay of correlations.*

5. SCALING BEHAVIOR.

5.1. The two-sites correlation function follows an asymptotic scaling law with exponent varying with the mutation probability. As shown by the following heuristic argument, the scaling exponent varies with the mutation probability in a piece-wise smooth way.

In the previous section we proved that

$$C_p(n) = \sum_{k=\lfloor n/2 \rfloor}^n C_p(k) (f(p) \nu_p(k, n) + g(p) \nu_p(k, n-1) + h(p) \nu_p(k, n-2)).$$

The distribution $k \mapsto \nu_p(k, n)$, which is unimodal with maximum at $k \approx n/(2-p)$, steepens around this maximum as n goes to infinity in such a way that

$$\sum_{k=\lfloor n/2 \rfloor}^n \nu_p(k, n) \approx \sum_{\ell(n) \leq k \leq u(n)} \nu_p(k, n),$$

where $\ell(n) < n/(2-p) < u(n)$ are such that $(2-p)\ell(n)/n, (2-p)u(n)/n \rightarrow 1$ as $n \rightarrow \infty$. Hence, assuming a slow variation in $k \mapsto C_p(k)$, we have

$$\begin{aligned} C_p(n) &\approx \sum_{\ell(n) \leq k \leq u(n)} C_p(k) (f(p) \nu_p(k, n) + g(p) \nu_p(k, n-1) + h(p) \nu_p(k, n-2)) \\ &\approx C\left(\frac{n}{2-p}\right) \sum_{\ell(n) \leq k \leq u(n)} (f(p) \nu_p(k, n) + g(p) \nu_p(k, n-1) + h(p) \nu_p(k, n-2)) \\ &\approx C\left(\frac{n}{2-p}\right) \sum_{k=\lfloor n/2 \rfloor}^n (f(p) \nu_p(k, n) + g(p) \nu_p(k, n-1) + h(p) \nu_p(k, n-2)) \\ &= C\left(\frac{n}{2-p}\right) (f(p) S_p(n) + g(p) S_p(n-1) + h(p) S_p(n-2)), \end{aligned}$$

where $S_n(p) := \sum_{k=\lfloor n/2 \rfloor}^n \nu_p(k, n) = (1 - (p-1)^n)/(2-p)$, as proved in Seccion 4. From this we finally obtain the approximate scaling relation

$$C_p((2-p)^k n_0) \approx \left(\frac{(1-2p)(2-3p)}{(2-p)} \right)^k C(n_0),$$

which traduces into the scaling law $C_p(n) \approx C_p(n_0) (n/n_0)^{-\beta_p}$ with

$$(10) \quad \beta_p := \frac{\log(2-p) - \log(1-2p) - \log(2-3p)}{\log(2-p)}.$$

From (4) we readily obtain the recurrence relation

$$C_p(n+1) = \frac{\sum_{k=\lfloor n/2 \rfloor}^{n-1} C_p(k) (f(p) \nu_p(k, n) + g(p) \nu_p(k, n-1) + h(p) \nu_p(k, n-2))}{1 + p^{n+1}(1-2p)}$$

which we use to numerically compute the two-point correlation function for different values of the mutation probability. As shown in Figure 1, the numerical computations confirm that the two-point correlation function approximately follows a power law behavior. Furthermore, according to Figure 2, the theoretically predicted exponents, $\{-\beta_p : 0 < p < 1\}$, fit very well the ones obtained by linear regression from the numerically computed correlation functions.

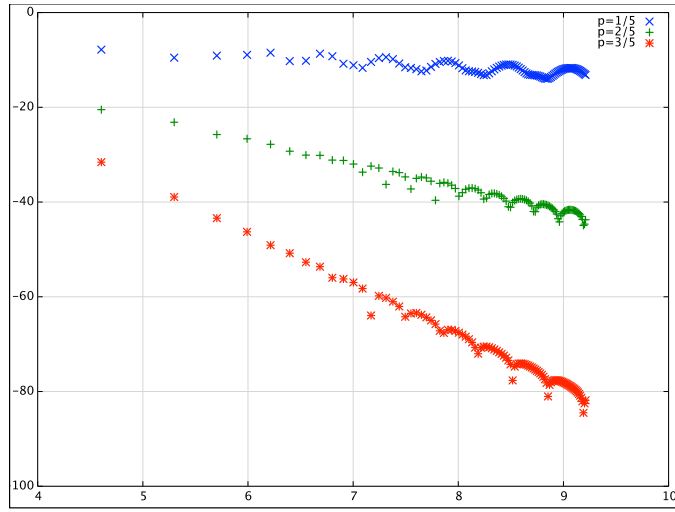


FIGURE 1. Log-log plot of the two-sites correlation function. A power law behavior clearly appears.

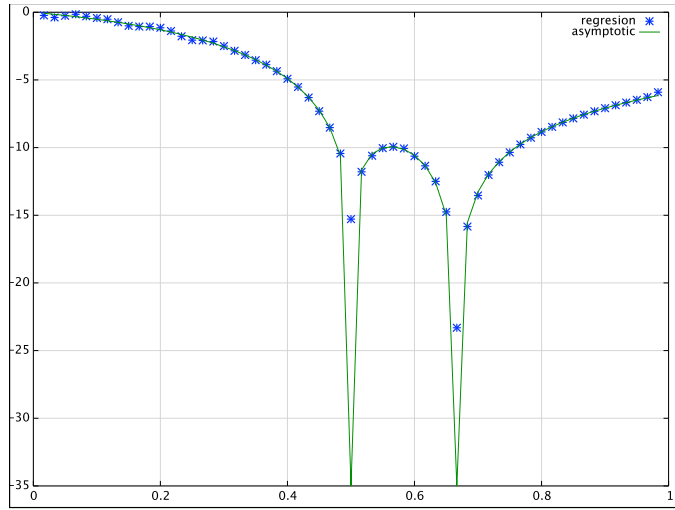


FIGURE 2. Exponents obtained by the best power law fit to the two-sites correlation function compared to the theoretical asymptotic exponents $-\beta_p$.

The heuristic argument developed above suggests that the stationary measure μ_p varies in a piecewise smooth way with p . This variation is reflected on the behavior of the two-sites correlation function C_p , which appears to follow a power law decay which prevails in the whole interval $0 < p < 1$, except for the two singularities we

find at $p = 1/2$ and $p = 2/3$. At precisely those values of p , the two-sites correlation function appears to decay faster than any power law.

5.2. The previous arguments can be refined into the following

Theorem 3. *Let $p \in (0, 1/2) \cup (2/3, 1)$ and suppose there are constants $0 < a \leq b$ and $n_0 \in \mathbb{N}$, such that*

$$(11) \quad n^{-b} \leq C_p(n) \leq n^{-a}$$

for all $n \geq n_0$. Then, there exist $m_0 \geq n_0$ and constants $A \leq 1 \leq B$ such that

$$A n^{-\beta p} \leq C_p(n) \leq B n^{-\beta p}$$

for all $n \geq m_0$.

Proof. Fix $\beta > b$ and let $d(x) := \sqrt{p(1-p)(2-p)(\beta+1)\log(x)/x}$. Following a standard concentration estimation, which we present in Appendix B, we prove that there exists $n_2 \in \mathbb{N}$ such that

$$(12) \quad \delta_n := n^b \left(\sum_{|n/k - (2-p)| > d(n)} \nu_p(k, n) \right) \leq n^{-(\beta-b)/2},$$

for all $n \geq n_2$.

Let $\ell, u : [1, \infty) \rightarrow [0, \infty)$ be such that

$$\ell(x) := \frac{x}{2-p+d(x)}, \quad u(x) := \frac{x}{2-p-d(x)},$$

with d as above. Let f_p, g_p, h_p and ν_p be as in Section 4, and define

$$\mathcal{W}_p(k, n) := f(p)\nu_p(k, n) + g(p)\nu_p(k, n-1) + h(p)\nu_p(k, n-2).$$

Since $\ell(n) \leq k \leq u(n) \Rightarrow |\frac{n}{k} - (2-p)| \leq d(n)$, $|C_p(k)| \leq 1$ for all k , and since $|f(p)| + |g(p)| + |h(p)| < 1$ for all $p \in (0, 1)$, then, using (4) and taking into account (12), we obtain

$$(13) \quad C_p(n) \leq \sum_{\ell(n) \leq k \leq u(n)} C_p(k) \mathcal{W}_p(k, n) + n^{-b} \delta_n$$

$$(14) \quad C_p(n) \geq \sum_{\ell(n) \leq k \leq u(n)} C_p(k) \mathcal{W}_p(k, n) - n^{-b} \delta_n$$

for each $n \geq n_2$.

In Appendix C we prove that there exists $n_1 \in \mathbb{N}$ such that $\mathcal{W}_p(k, n) > 0$ in the interval $\ell(n) \leq k \leq u(n)$ for all $p \in (0, 1/2) \cup (2/3, 1)$ and all $n \geq n_1$. For those values of p we can define a probability distribution $k \mapsto \mathbb{P}_p(k, n)$ proportional to $k \mapsto \mathcal{W}_p(k, n)$, in the interval $\ell(n) \leq k \leq u(n)$.

Assuming (11) we can rewrite Inequalities (13) and 14 as

$$\begin{aligned} C_p(n) &\leq \left(\sum_{k=\lfloor n/2 \rfloor}^n \mathcal{W}_p(n, k) + 2\delta_n \right) \mathbb{E}_{p,n}(C_p), \\ C_p(n) &\geq \left(\sum_{k=\lfloor n/2 \rfloor}^n \mathcal{W}_p(n, k) - 2\delta_n \right) \mathbb{E}_{p,n}(C_p), \end{aligned}$$

where $\mathbb{E}_{p,n}(C_p)$ denotes the mean value of C_p with respect to $\mathbb{P}_p(k, n)$. We have proved in Section 4, that $S_p(n) := \sum_{k=\lfloor n/2 \rfloor}^n \nu_p(k, n) = (2-p)^{-1}(1-(p-1)^n)$, therefore $\sum_{k=\lfloor n/2 \rfloor}^n \mathcal{W}_p(k, n) = (1-2p)(2-3p)/(2-p) - 2p(p-1)^n$ and we obtain

$$\begin{aligned} C_p(n) &\leq \left(\frac{(1-2p)(2-3p)}{2-p} + 3\delta_n \right) \mathbb{E}_{p,n}(C_p) \\ &\leq \left(\frac{(1-2p)(2-3p)}{2-p} + 3\delta_n \right) \max_{\ell(n) \leq k \leq u(n)} C_p(k), \\ C_p(n) &\geq \left(\frac{(1-2p)(2-3p)}{2-p} - 3\delta_n \right) \mathbb{E}_{p,n}(C_p) \\ &\geq \left(\frac{(1-2p)(2-3p)}{2-p} + 3\delta_n \right) \min_{\ell(n) \leq k \leq u(n)} C_p(k), \end{aligned}$$

for all $n \geq \max(n_1, n_2)$. To easy the notations, let

$$\lambda_p := (2-p), \quad \phi(x) := \ell(\lambda_p x)/\lambda_p \text{ and } \psi(x) := u(\lambda_p x)/\lambda_p.$$

With this we can rewrite the previous inequality as

$$\lambda_p^{-\beta_p - \eta_{\lambda_p x}} \min_{\lambda_p \phi(x) \leq y \leq \lambda_p \psi(x)} C_p([y]) \leq C_p([\lambda_p x]) \leq \lambda_p^{-\beta_p + \eta_{\lambda_p x}} \max_{\lambda_p \phi(x) \leq y \leq \lambda_p \psi(x)} C_p([y]),$$

with $\eta_x := 3\lambda_p^{-\beta_p} (x-1)^{-(\beta-b)/2} / \log(\lambda_p) \geq 3\lambda_p^{-\beta_p} \delta_{[x]} / \log(\lambda_p)$. It follows from this, by straightforward induction, that

$$\begin{aligned} C_p([\lambda_p^k x]) &\leq m_p^{-k\beta_p + \sum_{j=0}^{k-1} \eta_{\lambda_p \phi^j(\lambda_p^{k-1}x)}} \max_{\lambda_p \phi^k(\lambda_p^{k-1}x) \leq z \leq \lambda_p \psi^2(\lambda_p^{k-1}x)} C_p([z]), \\ C_p([\lambda_p^k x]) &\geq m_p^{-k\beta_p - \sum_{j=0}^{k-1} \eta_{\lambda_p \phi^j(\lambda_p^{k-1}x)}} \min_{\lambda_p \phi^k(\lambda_p^{k-1}x) \leq z \leq \lambda_p \psi^2(\lambda_p^{k-1}x)} C_p([z]), \end{aligned}$$

for all $k \in \mathbb{N}$ and $x \geq \max(n_1, n_2)$.

To finish the proof we use the following inequalities which we prove in Appendix D: There exists $x_0 > e$ such that for all $x \geq x_0$ there are constants $0 < Q(x) < 1 < R(x)$ such that for every $1 \leq j \leq k \in \mathbb{N}$ we have

$$Q(x) \lambda_p^{k-j} x \leq \lambda_p \phi^j(\lambda_p^{k-1} x) \leq u^j(\lambda_p^{k-1} x) \leq R(x) \lambda_p^{k-j} x.$$

Furthermore, $Q(x), R(x) \rightarrow 1$ when $x \rightarrow \infty$.

Using the above estimates and taking into account Hypothesis (11), we obtain, for all $p \in (0, 1/2) \cup (2/3, 1)$ and all $x \geq m_0 := \max(x_0, n_0, n_1, n_2)$, the inequalities

$$\begin{aligned} C_p([\lambda_p^k x]) &\leq \lambda_p^{-k\beta_p + \sum_{j=0}^{k-1} \eta_{Q(x) \lambda_p^{k-j} x}} (Q(x) x)^{-a} \\ &\leq \lambda_p^{-k\beta_p + \alpha_p(x)} (Q(x) x)^{-a}, \\ C_p([\lambda_p^k x]) &\geq \lambda_p^{-k\beta_p - \sum_{j=0}^{k-1} \eta_{Q(x) \lambda_p^{k-j} x}} (R(x) x)^{-b} \\ &\geq \lambda_p^{-k\beta_p - \alpha_p(x)} (R(x) x)^{-b}, \end{aligned}$$

where

$$\alpha_p(x) := \frac{3 \times (2\lambda_p)^{(\beta-b)/2}}{(Q(x) x)^{(\beta-b)/2} \lambda_p^{\beta_p} \log(\lambda_p) \left(\lambda_p^{(\beta-b)/2} - 1 \right)} \rightarrow 0 \text{ as } x \rightarrow \infty.$$

From these inequalities it follows that

$$\left((3/2)^{-\beta_p} R(x)^{-b} x^{\beta_p-b} \lambda_p^{-\alpha(x)} \right) n^{-\beta_p} \leq C_p(n) \leq \left(Q(x)^{-a} x^{\beta_p-a} \lambda_p^{-\alpha(x)} \right) n^{-\beta_p},$$

and the proof is finished. \square

In Appendix E we prove that for $p \leq 1/5$ there are constants $0 < a \leq b$ and $n_0 \in \mathbb{N}$, such that

$$n^{-b} \leq C_p(n) \leq n^{-a}$$

for all $n \geq n_0$. In this way we establish the asymptotic scaling behavior for small mutation probabilities.

6. CONCLUSIONS.

The expansion–modification system we have analyzed belongs to a class of stochastic dynamical systems defined by the action of a global random substitution. The existence of a unique stationary measure depends on the primitivity of the stochastic matrices describing the dynamics of the finite size marginals. The computations used to prove decay of correlations of the stationary measure, which rely on Equation (4), could be carried on in more general cases where relations similar to this one hold.

The asymptotic scaling behavior of the correlation function, which reflects the self-similar behavior (in a stochastic sense) of the system, is expected to take place in more general cases as well. It is important here to mention the work by Messer and co-authors [11], where a model generalizing Li’s is studied, and the work of Mansilla and Cocho [9] where the correlation function of Li’s model is studied. In the former the authors deduce an asymptotic scaling behavior from a closed expression for the correlations function and in the latter the authors obtain upper and lower bounds for a “dynamical” exponent of the correlation function. The present work follows similar ideas but in the framework of a rigorous study of the expansion–modification systems. We were able to prove the existence and uniqueness of the stationary measure, which attracts all initial distributions as times goes to infinity (Theorem 1). We also proved that this stationary measure exhibits decay of correlations (Theorem 2). We studied the scaling behavior of the correlation function and we deduced an expression (Equation (10)) for the scaling exponent as a function of the mutation probability. We rigorously established the validity of that expression in case of low mutation probabilities (Theorem 3).

The scaling behavior of the correlation function implies a scaling behavior for the so called power spectrum $f(\omega) := |\mathcal{F}(C_p)(\omega)|$, where $\mathcal{F}(C_p)$ denotes the discrete Fourier transform of the correlation function. A straightforward computation shows that $f(\omega) = \mathcal{O}(\omega^{-\alpha_p})$ with

$$\alpha_p := 1 - \beta_p = \frac{\log(2-p) - \log(1-2p) - \log(2-3p)}{\log(2-p)}.$$

It is worth to mention here the work by M. Zakz [16] where the power spectrum of systems similar to Li’s is studied. There, a scaling law is deduced from approximative recurrence relations for the Fourier transform of a realization of the system.

In a recent paper the expansion–modification system has been studied in relation to the universality of the rank–ordering distributions [1]. In that work the authors numerically found an order–disorder transition which would manifest itself on the scaling behavior. According to them, there would be a critical mutation probability $p_c \approx 0.4$, such that for $p > p_c$, long–range correlations and consequently the scaling behavior of C_p would disappear. As we have shown, this kind of order–disorder transition does not really occur. The apparent drop of long–range correlations for large p can be explained by the lack of statistics. Indeed, a huge amount of data is needed to empirically compute correlation functions with fast power law decay. In order to observe a power law decay with exponent -5 up to two decades, we would need of the order of 10^{10} sample sequences in $\{0,1\}^{10^2}$. These sample sequences should be generated by the action of the expansion–dynamics, over an arbitrary seed, after a sufficiently long transient. That explains why the scaling behavior of the C_p is very difficult to observe from empirical computations for $p \geq 0.4$, where $\beta_p > 5$ (see Figure 2).

REFERENCES

- [1] R. Alvarez–Martínez, G. Martínez-Mekler and G. Cocho “Order–disorder transition in conflicting dynamics leading to rank–frequency generalized beta distributions” *Physica A* **390** (1) 120–130 (2011).
- [2] A. L. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsu, C.-K. Peng, M. Simons, H. E. Stanley, “Long-range correlation properties of coding and noncoding DNA sequences: GenBank Analysis”, *Phys. Rev. E* **51**, 5084–5091 (1995).
- [3] W. Li, “Spatial 1/f spectra in open dynamical systems”, *Europhysics Letters* **10** (5) 395–400 (1989).
- [4] W. Li, “Expansion–modification systems: A model for spatial 1/f spectra”, *Physical Review A* **43** (10) 5240–5260 (1991).
- [5] W. Li and K. Kaneko, “Long-range correlation and partial 1/f a spectrum in a noncoding DNA sequence”, *Europhysics Letters* **17**, 655–660 (1992).
- [6] W. Li, “The study of correlation structures of DNA sequences: a critical review”, *Computers Chem.* **21**, 257–271 (1997).
- [7] P. Liò and N. Goldman, “Models of molecular evolution and phylogeny”, *Genome Research* **8**, 1233–1244 (1998).
- [8] J. Ma, A. Ratan, B. J. Raney, B. B. Suh, M. Miller, D. Haussler, “The infinite site model of genome evolution”, *Proc. Nat. Acad. Sci.* **105**, 14254–14261 (2008).
- [9] R. Mansilla and G. Cocho, “Multiscaling in expansion–modification systems: An explanation for long range correlation in DNA”, *Complex Systems* **12**, 207–240 (2000).
- [10] G. Martínez-Mekler, R. Alvarez–Martínez, M. Beltrán del Río, R. Mansilla, P. Miramontes, G. Cocho, “Universality of rank–ordering distributions in the arts and sciences”, *Plos One* **4**, e4791, 1–7 (2008).
- [11] Philipp W. Messer, Peter F. Arndt, and Michael Lässig, “Solvable sequence evolution models and genomic correlations”, *Phys. Rev. Lett.* **94**, 138103 (2005).
- [12] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, “Long-range correlations in nucleotide sequences”, *Nature* **356** 168–179 (1992).
- [13] D. B. Saakian, “Evolution models with base substitutions, insertions, deletions, and selection”, *Phys. Rev. E* **78**, 0611920 (2008).
- [14] M. Sobottka and A.G. Hart, “A model capturing novel strand symmetries in bacterial DNA”, *Biochem. & Biophys. Research Commun.* **410**, 823–828 (2011).
- [15] H. Robbins, “A Remark on Stirling’s Formula”, *The American Mathematical Monthly* **62** (1) 26–29 (1955).
- [16] M. Zaks, “Multifractal Fourier spectra and power law decay of correlations in random substitution sequences”, *Physical Review E* **65**, 011111 (2001).

APPENDIX A.

Claim 1. *For each $\mathbf{a}, \mathbf{b} \in \{0, 1\}^{\ell+1}$ there exists $\{\mathbf{s}^0, \mathbf{s}^1, \dots, \mathbf{s}^n\} \subset \{\mathbf{e}, \mathbf{m}\}^+$, such that $\mathbf{b} \sqsubseteq \mathbf{s}^n \circ \dots \circ \mathbf{s}^1 \circ \mathbf{s}^0(\mathbf{a})$.*

Proof. For $\ell \in \mathbb{N}_0$ and $\mathbf{a} \in \{0, 1\}^{\ell+1}$, let $\mathbf{s} = \mathbf{e}^{\ell+1}$ whenever $a_0 = 0$, otherwise let $\mathbf{s} = \mathbf{m}^{\ell}$. Clearly, for each $n > \lceil \log(\ell+1)/\log(2) \rceil$ we have

$$0^{\ell+1} \sqsubseteq \mathbf{t}^n \circ \dots \circ \mathbf{t}^1 \circ \mathbf{s}(\mathbf{a}),$$

where $\mathbf{t}^j \in \{\mathbf{e}, \mathbf{m}\}^+$ is such that $\mathbf{e}^{\ell+1} \sqsubseteq \mathbf{t}^j$ for each $1 \leq j \leq n$. We will prove that for each $\mathbf{b} \in \{0, 1\}^{\ell+1}$ there exists $\{\mathbf{s}^0, \mathbf{s}^1, \dots, \mathbf{s}^k\} \subset \{\mathbf{e}, \mathbf{m}\}^+$ such that

$$\mathbf{b} \sqsubseteq \mathbf{s}^k \circ \dots \circ \mathbf{s}^1 \circ \mathbf{s}^0(0^{\ell+1}),$$

which readily implies the claim.

Since $0 \sqsubseteq \mathbf{e}(0)$ and $1 \sqsubseteq \mathbf{m}(0)$, the claim holds for $\ell = 0$. Assuming the claim for $\ell = l-1$ we have, for all $\mathbf{c} \in \{0, 1\}^{l+1}$, a sequence $\{\mathbf{s}^0, \mathbf{s}^1, \dots, \mathbf{s}^k\} \subset \{\mathbf{e}, \mathbf{m}\}^+$ such that $\mathbf{c}_1^l \sqsubseteq \mathbf{s}^k \circ \dots \circ \mathbf{s}^1 \circ \mathbf{s}^0(0^l)$. If k is even and $c_0 = 1$ or k is odd and $c_0 = 0$, by taking $\mathbf{t}^j := \mathbf{m}^j$ for $0 \leq j \leq k$, we have

$$\mathbf{c} = c_0 \mathbf{c}_1^m \sqsubseteq \mathbf{m}^{k+1}(0) \mathbf{s}^k \circ \dots \circ \mathbf{s}^1 \circ \mathbf{s}^0(0^{l+1}) = \mathbf{t}^k \circ \dots \circ \mathbf{t}^1 \circ \mathbf{t}^0(0^{l+1}).$$

On the other hand, if k is even and $c_0 = 0$ or k is odd and $c_0 = 1$, by taking $\mathbf{t}^{j+1} := \mathbf{m}^j$ for $0 \leq j \leq k$, and $\mathbf{t}^0 := \mathbf{e}^{l+1}$, we have

$$\mathbf{c} = c_0 \mathbf{c}_1^m \sqsubseteq \mathbf{m}^{k+2}(0) \mathbf{s}^k \circ \dots \circ \mathbf{s}^1 \circ \mathbf{s}^0(0^{l+1}) = \mathbf{t}^{k+1} \circ \dots \circ \mathbf{t}^1 \circ \mathbf{t}^0(0^{l+1}),$$

and the claim follows. \square

APPENDIX B.

We start with the following.

Lemma 1. *For each $p \in (0, 1)$ the function*

$$q \mapsto I_p(q) := q/(q+1) \log(q/(1-p)) + (1-q)/(q+1) \log((1-q)/p)$$

is non negative, strictly convex, and satisfies

$$\min\{I_p(q) : q \in (0, 1)\} = I_p(1-p) \equiv 0.$$

Proof. Since $x \mapsto -\log(x)$ is a concave function, then

$$\begin{aligned} I_p(q) &= \frac{1}{q+1} \left(-q \log\left(\frac{1-p}{q}\right) - (1-q) \log\left(\frac{p}{1-q}\right) \right) \\ &\geq \frac{1}{q+1} \left(-\log\left(q \frac{1-p}{q} + (1-q) \frac{p}{1-q}\right) \right) = 0. \end{aligned}$$

On the other hand,

$$\frac{dI_p(q)}{dq} = \frac{1}{(q+1)^2} \left(\log\left(\frac{q}{1-p}\right) - 2 \log\left(\frac{1-q}{p}\right) \right) = 0 \Leftrightarrow q = 1-p.$$

In this way we prove that I_p is non-negative with minimum at $q = 1-p$. Now,

$$\frac{d^2 I_p(q)}{dq^2} = \frac{1}{q(1-q^2)} + \frac{2}{(q+1)^3} \left(2 \log\left(\frac{1-q}{p}\right) + \left(\frac{1-p}{q}\right) \right) > 0$$

for all $p, q \in (0, 1)$. For this, note that if $1 - q \geq p$ then $1 - p \geq q$ and in this case we have

$$\frac{d^2 I_p(q)}{dq^2} \geq \frac{1}{q(1 - q^2)} > 0,$$

otherwise, for $1 - q < p$ then $1 - p < q$, and taking into account that $-\log(x) \geq 1 - x$, we obtain

$$\begin{aligned} \frac{d^2 I_p(q)}{dq^2} &= \frac{1}{q(1 - q^2)} - \frac{2}{(q + 1)^3} \log \left(\left(\frac{(1 - q)^2}{p} \right) \frac{(1 - p)}{q} \right) \\ &\geq \frac{1}{q(1 - q^2)} + \frac{2}{(q + 1)^3} \left(1 - \left(\frac{(1 - q)^2}{p} \right) \frac{(1 - p)}{q} \right) \\ &\geq \frac{1}{q(q + 1)} \left(\frac{1}{1 - q} - 2 \frac{(1 - q)q}{(q + 1)^2} \right) = \frac{1 + 3q^2}{q(q + 1)^3(1 - q)} > 0. \end{aligned}$$

Therefore I_p is strictly convex. \square

Let $0 < \beta < b$ be given, and let $d(n)$ be defined as in Subsection 5.2. We have the following:

Claim 2. *For each $p \in (0, 1)$, there exists $n_2 \in \mathbb{N}$ such that*

$$\delta_n := n^b \left(\sum_{|n/k - (2-p)| > d(n)} \nu_p(k, n) \right) \leq n^{-(\beta-b)/2},$$

for all $n \geq n_2$.

Proof. A very useful refinement of Stirling's approximation, first published in [15], states that

$$\sqrt{2\pi n} n^n \exp \left(-n + \frac{1}{12n + 1} \right) \leq n! \leq \sqrt{2\pi n} n^n \exp \left(-n + \frac{1}{12n} \right)$$

for all $n \in \mathbb{N}$. Hence, for each $p \in (0, 1)$, $n \geq 3$, and $[n/2] < k < n$, we have

$$\exp(-\epsilon_{n,k}) \leq \frac{\sqrt{2\pi k (n/k - 1) (2 - n/k)}}{k^k (n - k)^{-(n-k)} (2k - n)^{-(2k-n)}} \times \binom{k}{n - k} \leq \exp(+\epsilon_{n,k}),$$

with $\epsilon_{n,k} = (4 \min(n - k, 2k - n))^{-1}$. A simple computation shows that

$$k^k (n - k)^{-(n-k)} (2k - n)^{-(2k-n)} (1 - p)^{n-k} p^{2k-n} = \exp(-n I_p(1 - k/n)),$$

with $I_p(q) := q/(q + 1) \log(q/(1 - p)) + (1 - q)/(q + 1) \log((1 - q)/p)$ for each $q \in (0, 1)$. Hence

$$(15) \quad \frac{e^{-n I_p(n/k - 1) - \epsilon_{n,k}}}{\sqrt{2\pi k (n/k - 1) (2 - n/k)}} \leq \nu_p(k + 1, n + 1) \leq \frac{e^{-n I_p(n/k - 1) + \epsilon_{n,k}}}{\sqrt{2\pi k (n/k - 1) (2 - n/k)}},$$

for each $n \geq 3$ and $[n/2] \leq k \leq n$. On the other hand,

$$\begin{aligned} \nu_p(n/2 + 1, n + 1) &= (1 - p)^{n/2} \equiv \lim_{q \rightarrow 1} \exp(-n I_p(q)), \\ \nu_p(n + 1, n + 1) &= p^n \equiv \lim_{q \rightarrow 0} \exp(-n I_p(q)), \end{aligned}$$

hence, by using

$$A^\pm := \begin{cases} \frac{\exp(\pm 1/(4 \min(n - k, 2k - n)))}{\sqrt{2\pi k (n/k - 1) (2 - n/k)}} & \text{if } n/2 < k < n, \\ 1 & \text{otherwise,} \end{cases}$$

we can extend (15) to

$$(16) \quad e^{-n I_p(n/k-1)} A^- \leq \nu_p(k+1, n+1) \leq e^{-n I_p(n/k-1)} A^+$$

which holds for all $n \geq 3$ and $[n/2] \leq k \leq n$. A simple computation shows that

$$A^+ \leq \max \left(1, \frac{e^{1/4} \sqrt{n-1}}{\sqrt{2\pi(n-2)}}, \frac{e^{1/4} \sqrt{n+2}}{\sqrt{4\pi(n-2)}} \right) = 1,$$

for $n \geq 3$.

We have already proved that the function $q \mapsto I_p(q)$ is non negative, vanishes only at $q = 1 - p$, and is strictly convex. Hence, Taylor's Theorem ensures that

$$(q - (1 - p))^2 \times \min_{|q - (1 - p)| \leq \epsilon} \left| \frac{d^2 I_p}{dq^2} \right|_q \leq I_q(q) \leq (q - (1 - p))^2 \times \max_{|q - (1 - p)| \leq \epsilon} \left| \frac{d^2 I_p}{dq^2} \right|_q.$$

Since $d^2 I_p/dq^2|_{q=1-p} = (p(1-p)(2-p))^{-1}$ and $q \mapsto d^2 I_p/dq^2$ is continuous, then, for all $\alpha \in (0, 1)$ there exists $\epsilon_\alpha > 0$ such that

$$\frac{\alpha (q - (1 - p))^2}{p(1 - p)(2 - p)} \leq I_q(q) \leq \frac{\alpha^{-1} (q - (1 - p))^2}{p(1 - p)(2 - p)}$$

for all $q \in (1 - p - \epsilon_\alpha, 1 - p + \epsilon_\alpha)$. With this, and using (16), we obtain

$$0 \leq \sum_{|k/n - (2-p)| \geq \epsilon} \nu_p(k+1, n+1) \leq n \times \exp(-n I_p(\epsilon)) \leq n \times \exp \left(\frac{-n \alpha \epsilon_\alpha^2}{p(1 - p)(2 - p)} \right)$$

for all $\epsilon \leq \epsilon_\alpha$. By taking n such that $d(n) := \sqrt{p(1 - p)(2 - p)(\beta + 1) \log(n)/n} \leq \epsilon_\alpha$ we obtain

$$0 \leq \sum_{|k/n - (2-p)| \geq d(n)} \nu_p(k+1, n+1) \leq n^{1 - \alpha(\beta + 1)},$$

and the claim follows with $\alpha = (b + \beta + 2)/(2\beta + 2)$ and n_2 such that $d(n) \leq \epsilon_\alpha$ for all $n \geq n_2$. \square

APPENDIX C.

Claim 3. *For each $p \in (0, 1/2) \cup (2/3, 1)$ there exists $n_1 \in \mathbb{N}$ such that $\mathcal{W}_p(k, n) > 0$ in the interval $\ell(n) \leq k \leq u(n)$ for all $p \in (0, 1/2) \cup (2/3, 1)$ and all $n \geq n_1$.*

Proof. A simple computation shows that

$$\begin{aligned} \mathcal{W}_p(k, n) &= \frac{(1 - p)^{n-k} p^{2k-n} (k-1)!}{(n-k)!(2k-n+1)!} Q_p(n, k), \text{ with} \\ Q_p(k, n) &:= ((1 - 2p)(2n - 3k)(2k - n + 1) + p(2n - 3k - 2)(n - k)) \end{aligned}$$

If $p \in (0, 1/2)$ and n is so large that $p \leq 1/2 - d(n/2)$, then $3u(n/(2-p)) \leq 2n - 3$ and in this case $Q_p(k, n) > 0$ for all $\ell(n/(2-p)) \leq k \leq u(n/(2-p))$. It is easy to check that

$$\frac{1 - p - d(n/2)}{(p + d(n/2))(n + 1)} \leq \frac{n - k}{2k - n + 1} \leq \frac{1 - p + d(n/2)}{p - d(n/2)}$$

for $\ell(n/(2-p)) \leq k \leq u(n/(2-p))$. Hence, if $p \in (1/2, 1)$ and n is so large that $p \geq 1/2 + d(n/2) + 1/\sqrt{n}$, then $3\ell(n/(2-p)) \geq 2n + \sqrt{n}$, and in this case $Q_p(k, n)$ have the same sign as

$$1 - \frac{p}{2p-1} \left(1 + \frac{2}{3k-2n} \right) \frac{n-k}{2k-n+1}$$

for all $\ell(n/(2-p)) \leq k \leq u(n/(2-p))$. Here we have two possibilities, either $p < 2/3$ and

$$\lim_{n \rightarrow \infty} \frac{p}{2p-1} \left(\left(1 + \frac{2}{n} \right) \frac{1-p+d(n/2)}{p+d(n/2)} - 1 \right) = \frac{1-p}{2-p} < 1,$$

or $p > 2/3$ and then

$$\lim_{n \rightarrow \infty} \frac{p}{2p-1} \left(\left(1 + \frac{2}{\sqrt{n}} \right) \frac{1-p-d(n/2)}{(p+d(n/2))(n+1)} - 1 \right) = \frac{1-p}{2-p} < 1.$$

From all this we conclude that the sign of $\mathcal{W}_p(k, n)$ remains constant in the interval $\ell(n/(2-p)) \leq k \leq u(n/(2-p))$ for all $p \in (0, 1) \setminus \{1/2, 2/3\}$ and all sufficiently large n . Furthermore, this sign is positive for $p \in (0, 1/2) \cup (2/3, 1)$ and negative in the interval $(1/2, 2/3)$. \square

APPENDIX D.

Claim 4. *There exists $x_0 \geq e$ such that, for each $x \geq x_0$ there are constants $0 < Q(x) < 1 < R(x)$ such that for every $1 \leq j \leq k \in \mathbb{N}$ we have*

$$Q(x) \lambda_p^{k-j} x \leq \lambda_p \phi^j(\lambda^{k-1} x) \leq u^j(\lambda_p^{k-1} x) \leq R(x) \lambda_p^{k-j} x.$$

Furthermore, $Q(x), R(x) \rightarrow 1$ when $x \rightarrow \infty$.

Proof. A straightforward computation shows that

$$\lambda_p \phi^j(\lambda_p^{k-1} x) = \ell(\lambda_p^k x) \text{ and } \lambda_p \psi^j(\lambda_p^{k-1} x) = u(\lambda_p^k x),$$

for all $x \geq e$, $k \in \mathbb{N}$ and $1 \leq j \leq k$. It is easily checked that, for each $p \in (0, 1)$ and $\beta > b$, there exists $x_1 \geq e$ such that both ℓ and u are increasing functions in $[x_0, \infty)$.

For each $p \in (0, 1)$ and $x \geq x_1$ let $Q(x)$ be the largest solution to

$$Q(x) = \exp \left(- \frac{d(x)}{\sqrt{Q(x)} \lambda_p} \sum_{m=0}^{\infty} \sqrt{\frac{m+1}{\lambda_p^m}} \right).$$

It is not difficult to check that $Q(x) \in (0, 1)$ and since $d(x) \rightarrow 0$ as $x \rightarrow \infty$, then $Q(x) \rightarrow 1$ as $x \rightarrow \infty$.

Now, fix $k \in \mathbb{N}$, and $x_0 \geq x_1$ large enough so that $\lambda_p Q(x) \geq 1$ for all $x \geq x_0$. Let $Q_{k,0}(x) := 1$, and define recursively

$$Q_{k,j+1}(x) := \frac{Q_{k,j}(x)}{1 + \lambda_p^{-(k-j)/2-1} Q(x)^{-1/2} \sqrt{k-j+1} d(x)}$$

for $0 \leq j \leq k-1$. Clearly

$$\begin{aligned} 1 \geq Q_{k,j}(x) &= \prod_{i=0}^{j-1} \left(1 + \lambda_p^{-(k-i)/2-1} Q(x)^{-1/2} \sqrt{k-i+1} d(x) \right)^{-1} \\ &\geq \exp \left(-\frac{d(x)}{\sqrt{Q(x)}} \sum_{i=0}^{j-1} \sqrt{k-i+1} \lambda_p^{-(k-i)/2-1} \right) \\ &\geq \exp \left(-\frac{d(x)}{\sqrt{Q(x)} \lambda_p} \sum_{m=0}^{\infty} \sqrt{\frac{m+1}{\lambda_p^m}} \right) = Q(x). \end{aligned}$$

Since $x \geq e > \lambda_p$, then $x^{k+1} \geq \lambda_p^k x$, and therefore

$$\begin{aligned} d(\lambda_p^k x) &= \sqrt{\frac{p(1-p)(2-p)(\beta+1) \log(\lambda_p^k x)}{\lambda_p^k x}} \\ &\leq \lambda_p^{-k/2} \sqrt{\frac{p(1-p)(2-p)(\beta+1)(k+1) \log(x)}{x}} = \sqrt{\frac{k+1}{\lambda_p^k}} d(x). \end{aligned}$$

Hence, for $j=1$ we have

$$\ell(\lambda_p^k x) = \frac{\lambda_p^{k-1} x}{1 + d(\lambda_p^k x)/\lambda_p} \geq \frac{\lambda_p^{k-1} x}{1 + \lambda_p^{-k/2-1} \sqrt{k+1} d(x)} = Q_{k,1}(x) \lambda_p^{k-1} x.$$

Suppose that $\ell^j(\lambda_p^k x) \geq Q_{k,j}(x) \lambda_p^{k-j} x$ for $j < k$. Since $Q_{k,j}(x) \lambda_p^{k-j} x \leq \lambda_p^{k-j} x \leq x^{k-j+1}$, then

$$\begin{aligned} d(Q_{k,j} \lambda_p^{k-j} x) &= \sqrt{\frac{p(1-p)(2-p)(\beta+1) \log(Q_{k,j}(x) \lambda_p^{k-j} x)}{\lambda_p^{k-j} Q_{k,j}(x) x}} \\ &\leq \lambda_p^{-(k-j)/2} Q_{k,j}(x)^{-1/2} \sqrt{k-j+1} d(x) \\ &\leq \lambda_p^{-(k-j)/2} Q(x)^{-1/2} \sqrt{k-j+1} d(x). \end{aligned}$$

Taking into account that $y \mapsto \ell(y)$ is an increasing function for $y \geq x_0$, and since $Q_{k,j}(x) \lambda_p^{k-j} x \geq Q(x) x \geq x_0$, then

$$\begin{aligned} \ell^{j+1}(\lambda_p^k x) &\geq \frac{Q_{k,j}(x) \lambda_p^{k-j} x}{\lambda_p + d(Q_{k,j}(x) \lambda_p^{k-j} x)} \\ &\geq \frac{Q_j(x) \lambda_p^{k-j-1} x}{1 + \lambda_p^{(k-j)/2-1} Q(x)^{-1/2} \sqrt{k-j+1} d(x)} \\ &= Q_{k,j+1}(x) \lambda_p^{k-j-1} x. \end{aligned}$$

In this way we have proved that

$$\ell^j(\lambda_p^k x) \geq Q_{k,j}(x) \lambda_p^{k-j} x \geq Q(x) \lambda_p^{k-j} x,$$

for all $k \in \mathbb{N}$ and $1 \leq j \leq k$, and $x \geq x_0$.

By taking $R(x)$ the smallest solution to

$$R(x) = \exp \left(\frac{d(x)}{\sqrt{R(x)} \lambda_p} \sum_{m=0}^{\infty} \sqrt{\frac{m+1}{\lambda_p^m}} \right),$$

the previous argument can be easily adapted to deduce

$$u^j(\lambda_p^k x) \leq R(x) \lambda_p^{k-j} x,$$

for all $k \in \mathbb{N}$ and $1 \leq j \leq k$, and $x \geq x_0$. \square

APPENDIX E.

Claim 5. *For each $p \in (0, 1/10]$ there are constants $0 < a \leq b$ and $n_0 \in \mathbb{N}$, such that*

$$n^{-b} \leq C_p(n) \leq n^{-a}$$

for all $n \geq n_0$. In this way we establish the asymptotic scaling behavior in the interval $(0, 1/10]$.

Proof. Equation (4) can be rewritten as

$$(17) \quad C_p(n) = \frac{1}{1 + p^n(1 - 2p)} \sum_{k=\lfloor n/2 \rfloor}^{n-1} C_p(k) \mathcal{W}_p(k, n),$$

with $\mathcal{W}_p(k, n)$ as Appendix C. Since $p < 1/2$, the, as we proved in Appendix C, $\mathcal{W}_p(k, n) > 0$ whenever $3k \leq 2n - 3$. On the other hand, $q \mapsto I_p(q)$ is monotonously decreasing in $[0, 1 - p]$ and monotonously decreasing in $[1 - p, 1]$. Therefore, following the computations of Appendix B, we obtain

$$\begin{aligned} C_p(n) &\leq \frac{1}{1 + p^n(1 - 2p)} \left(\max_{\lfloor n/2 \rfloor \leq k \leq 2n/3} C_p(k) \right) \sum_{k < 2n/3} \mathcal{W}_p(k, n) \\ &\leq \frac{\max_{\lfloor n/2 \rfloor \leq k \leq 2n/3} C_p(k)}{1 + p^n(1 - 2p)} \left(\sum_{k=\lfloor n/2 \rfloor}^n \mathcal{W}_p(k, n) + \frac{n}{3} e^{-(n-1)I_p(\frac{1}{2} - \frac{3}{2n})} \right), \end{aligned}$$

for all $n \geq 4$. Now, since $1/2 - 3/2n < 3/4 \leq 1 - p$, and $q \mapsto I_p(q)$ is monotonously decreasing in $[0, 1 - p]$, then

$$I_p\left(\frac{1}{2} - \frac{3}{2n}\right) > I_p(1/2) = \frac{1}{3} \log\left(\frac{1}{4p(1-p)}\right)$$

for each $n \geq 4$ and $p \in (0, 1/10]$. With this, and taking into account that $\sum_k \mathcal{W}_p(k, n) = (1 - 2p)(2 - 3p)/(2 - p) - 2p(p - 1)^n$, we obtain

$$C_p(n) \leq \frac{\max_{k \leq 2n/3} C_p(k)}{1 + p^n(1 - 2p)} \left(\frac{(1 - 2p)(2 - 3p)}{2 - p} + 2p(1 - p)^n + \frac{n}{3} (4p(1 - p))^{(n-1)/3} \right)$$

for all $p \leq 1/10$ and $n \geq 4$. Hence, for $N \geq 4$, taking into account that $C_p(k) \leq 1/4$ for all $k \in \mathbb{N}$, we have

$$\begin{aligned} C_p(N + m) &\leq \frac{1}{4} \frac{(1 - 2p)(2 - 3p)}{2 - p} \exp(\gamma_p(N)) \text{ for each } 0 \leq m < N, \\ C_p(2N) &\leq \frac{1}{4} \left(\frac{(1 - 2p)(2 - 3p)}{2 - p} \right)^2 \exp(\gamma_p(N) + \gamma(2N)), \end{aligned}$$

where

$$\gamma_p(N) := \max_{N \leq n \leq 2N-1} \frac{1}{1 + p^n(1 - 2p)} \left(\frac{(2 - p)n(4p(1 - p))^{(n-1)/3}}{3(1 - 2p)(2 - 3p)} + \frac{2p(2 - p)(1 - p)^n}{(1 - 2p)(2 - 3p)} \right)$$

A simple recursion on these inequalities leads to

$$C_p(2^\kappa N + m) \leq \frac{1}{4} \left(\frac{(1-2p)(2-3p)}{2-p} \right)^\kappa \exp \left(\sum_{k=0}^{\kappa} \gamma_p(2^k N) \right),$$

for all $\kappa \in \mathbb{N}$ and each $0 \leq m < 2^\kappa N$. Since $\sum_{k=0}^{\infty} \gamma_p(2^k N) < \infty$ for all $p \in [0, 1]$, then we can find, for $\epsilon > 0$, an integer $N \geq 4$ such that

$$(18) \quad C_p(n) \leq n^{-((\log(2-p) - \log(1-2p) - \log(2-3p))/\log(2) + \epsilon)}$$

for each $n \geq N$.

The proof of the lower bound reduces to a recursion as well, but in order to establish the seed of the recursion, we will need a bit of brute force. Using (17) we explicitly compute the correlation function $C_p(n)$, for n small ². It can be checked that all of the functions $p \mapsto C_p(n)$ are positive, monotonously decreasing and log concave in $[0, 1/10]$. By log concave we mean that

$$\begin{aligned} \log(C_p(n)) &\geq (1-10p) \log(C_0(n)) + 10p \log(C_{1/5}(n)) \\ &= \log(C_0(n)) + 10p \log \left(\frac{C_{1/10}(n)}{C_0(n)} \right) \\ &= -\log(4) + 10p \log(4 C_{1/10}(n)), \end{aligned}$$

which holds for all $1 \leq n \leq 25$ and $p \in [0, 1/10]$. Here we are using that $C_0(n) = 1/4$ for all $n \in \mathbb{N}$. From the log concavity we can easily find power law bounding $C_p(n)$ from below. Indeed, by taking

$$b_p := 0.60206 + 5.62823p > \max_{12 \leq n \leq 25} \frac{\log(4) - 10p \log(4 C_{1/10}(n))}{\log(n)},$$

we have $C_p(n) \geq n^{-b(p)}$ for all $12 \leq n \leq 25$ and $p \in [0, 1/10]$.

As we showed above, $I_p(1/2 - 3/(2n)) > -\log(4p(1-p))/3$ for each $n \geq 4$ and $p \in (0, 1/10]$, therefore

$$\left| \sum_{k > 2n/3} \mathcal{W}_p(k, n) C_p(k) \right| \leq \frac{1}{4} \frac{n (4p(1-p))^{(n-1)/3}}{3}$$

for each $n \geq 4$ and $p \in [0, 1/10]$. Taking into account the power law bound obtained by direct computations, $C_p(n) \geq n^{-b(p)}$, and using (17), we have

$$\begin{aligned} C_p(n) &\geq \frac{\left(\sum_{k < 2n/3} \mathcal{W}_p(k, n) \right) \left(\min_{k < 2n/3} C_p(k) \right) - \frac{n}{12} (4p(1-p))^{(n-1)/3}}{1 + p^{n+1}(1-2p)} \\ &\geq \frac{\left(\sum_{k \leq n} \mathcal{W}_p(k, n) \right) \left(\frac{2n}{3} \right)^{-b_p} - \frac{n}{12} (4p(1-p))^{(n-1)/3}}{1 + p^{n+1}(1-2p)} \\ &\geq \frac{\left(\frac{(1-2p)(2-3p)}{2-p} - 2p(p-1)^n - 2^{b_p-2} \left(\frac{n}{3} \right)^{1+b_p} (4p(1-p))^{(n-1)/3} \right)}{1 + p^{n+1}(1-2p)} \left(\frac{2n}{3} \right)^{-b_p} \end{aligned}$$

²We used *Maxima* to compute $C_p(n)$ for $1 \leq n \leq 25$.

for all n such that $25 \leq n$ and $2n/3 \leq 25$, *i.e.*, $25 \leq n \leq 37$. It can be checked that

$$\alpha_p(n) := \frac{\left(\frac{(1-2p)(2-3p)}{2-p} - 2p(p-1)^n - 2^{b_p-2} \left(\frac{n}{3}\right)^{1+b_p} (4p(1-p))^{(n-1)/3}\right)}{1 + p^{n+1}(1-2p)} \left(\frac{3}{2}\right)^{b_p},$$

decreases with p and increases with n . Since $\alpha_{1/10}(25) \approx 1.0999111 > 1$, it follows that $C_p(n) \geq n^{-b_p}$ for all $12 \leq n \leq 37$ and $p \in [0, 1/10]$. From here, a standard induction implies that $C_p(n) \geq n^{-b_p}$ for all $n \geq 12$ and $p \in [0, 1/10]$, and the claim follows. \square

We could, in the previous claim, enlarge the interval of mutation probabilities in which upper and lower power law bounds for the correlation function can be found. To this aim, and following the same scheme of proof, it would be necessary to explicitly compute $C_p(n)$ for larger values of n . By doing so we could in principle replace $[0, 1/10]$ to $[0, p^*)$, where $p^* := \sup\{p \in (0, 1) : C_p(n) > 0 \forall n \in \mathbb{N}\} \approx 0.28$.

FACULTAD DE CIENCIAS, UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS, AVENIDA UNIVERSIDAD 1001, COLONIA CHAMILPA, CUERNAVACA C.P. 62209, MORELOS, MEXICO.

INSTITUTO DE FÍSICA, UNIVERSIDAD AUTÓNOMA DE SAN LUIS POTOSÍ, AVENIDA MANUEL NAVA 6, ZONA UNIVERSITARIA, 78290 SAN LUIS POTOSÍ, MÉXICO.